DOI: https://doi.org/10.24113/ijoscience.v4i3.124

# Correlation Enhanced Machine Learning Approach based Online News Popularity Prediction

Akanksha Kathal
M.Tech Scholar
SIRT
Bhopal, M. P, India
akanksha.kathal@gmail.com

Mayank Namdev
Professor
SIRT
Bhopal, M, P, India
mayank.namdev@gmail.com

***Abstract*: News popularity is the maximum growth of attention given for particular news article. The popularity of online news depends on various factors such as the number of social media, the number of visitor comments, the number of Likes, etc. It is therefore necessary to build an automatic decision support system to predict the popularity of the news as it will help in business intelligence too. The work presented in this study aims to find the best model to predict the popularity of online news using machine learning methods. Initially, correlation techniques are used to gain dependence on the popularity received from an article and to obtain attributes or characteristics that are optimal for subsequent classification. Data has been procured from UCI Machine Learning Repository with 39644 articles with sixty condition attributes and one decision attribute. Then different learning algorithms such as Proposed Hybrid SVM-RF, AdaBoost, LPBoost, and KNN are implemented in order to predict the news popularity. The performance of system is tested on the dataset which comes from UCI machine learning repository. The prediction performances of all methodologies are studied by considering evaluation measures. Hybrid SVM-RF turns out to be the best model for prediction and it has achieved accuracy of 99.6% for binary classification. Further this work is enhanced for multiclass classification with different learning algorithms such as Proposed Hybrid SVM-RF, Naïve Bayes and KNN. Hybrid SVM-RF had achieved the accuracy of about 73% accuracy as compared with other classifiers.**

**Keywords** – Machine Learning, Classification, Popularity Prediction, Correlation Co-efficient, Accuracy.

## 1. INTRODUCTION

In the digital world, online news is primary source of information [1]. Various businesses are keen to know what will be the future demand of online visitors. Popularity prediction is useful in many applications like media advertising, estimation of movie revenue, traffic management, economic trends forecasting. Popularity prediction is hard to capture as it depends upon various factors like its topic, text, timing, article's position on the web page, language, similarity with world's event, same subject historical popularity, time from news publication, season of article popularity, relevance to the physical world popular events [2,3].

The large numbers of prediction methods for various types of web content are proposed in the research of latest years [4]. This research work is focused on prediction task.

Number of shares is one of the factors to determine popularity of news articles. In this work, it is intended to find the better model to predict the online popularity of news by using different machine learning techniques [5].

Prediction of the popularity is considered into 2 parts i.e. popularity prediction before publication of news and popularity publication after publication of news [6]. Mostly popularity prediction before news publication is considered for study. In the latest years, different types of prediction methods for different types of web information have been proposed [7]. This study focused on prediction of large visitors' attention to particular news articles, its reasons, evaluated methodologies, considered parameters and improved results.

In the current scenario, this paper have proposed methodologies which provide a way to predict whether an article will become popular or not. The objective of the paper is to maximize the rate of prediction of the article by minimizing and selecting the optimum features [8-10]. Publishers can benefit by estimating the popularity of the

news content and strategize accordingly by focusing on the features obtained as a result of this analysis [11]. Further this paper is enhanced to make comparative analysis of multiclass (popular, Unpopular and Average) popularity prediction methods by considering parameters [12]. To fulfill the objectives of this paper, dataset of 39,797 news articles are collected from UCI machine learning repository which is a collection of Mashable's online news website [13]. Different machine learning algorithms are planned to implement on the dataset to evaluate and compare their performances.

## 2. RELATED WORK

The research work done so far mainly focused on the attributes of online content for estimating future popularity. Ilias_N._Lymperopoulos [1] predicts popularity of online content through forward extrapolation. Prediction results are improved as compared to existing methods and in terms of precision and accuracy.

Kelwin Fernandes [2] proposed online news popularity prediction using machine learning approaches such as RF, AdaBoost, SVM, KNN and Naïve bayes. Out of all classifiers RF is the better model showing 67% of accuracy.

He Ren and Quan Yang [3]find out the Mutual Information and Fisher Criterion to get maximum accuracy for feature selection. Random forest is used for classification and obtained an accuracy of 69%.

Sitaram [4] predicted popularity on twitter using Bagging, J48 Decision Trees, SVM and Naïve Bayes classifiers and achieved an accuracy of 84%.

Ioannis_Arapakis [5]categorized the features into 10 main headings. Articles are classiifed using Naïve Bayesian, Bagging, decision trees (J48), SVM and achieved the accuracy of 79.7%.

R. Shreyas [6] used the Random Forest approach to predict popular/unpopular articles and achieved the accuracy of 88.8%.

Swati Choudhary [7] used genetic algorithm to get the optimum attributes and further classified the data using different classifiers and obtained the highest accuracy of 91.96% with naïve bayes classifier.

## 3. PROPOSED WORK

In the current scenario, an algorithm is proposed which provide a way to predict whether an article will become popular or not. Figure 1 shows the overall architecture for prediction of popularity of online published news. The proposed work is presented for two cases, i.e. case I(for binary classification) and Case II (for multi-classification). The proposed work is designed for optimized feature selection for online news popularity prediction process.

Following diagram describes flow of News Popularity Prediction System

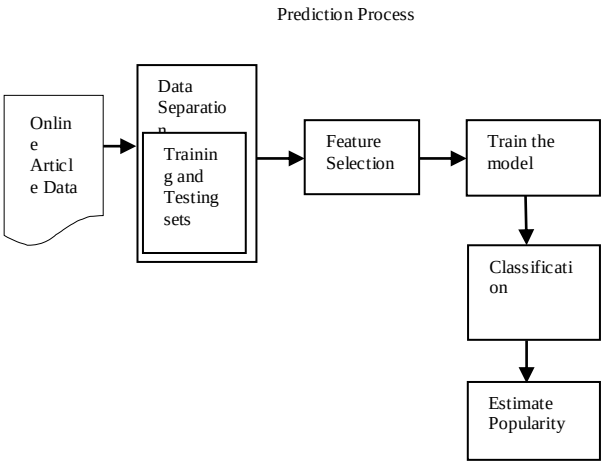


**Fig. 1: Flow diagram of News Popularity Prediction System**

### 3.1 Correlation Analysis

A bivariate analysis used for measuring the degree of association amongst two vectors say A and B is known as Correlation. In data mining, the value obtained after doing Correlation analysis varies between ±1. When this value is greater than 0, then a positive correlation exists and if this value is less than zero, then a negative correlation exists. If the value is 0, then the relationship between them is weak. For the proposed work that correlation value is selected whose value is positive one.

In this paper for feature selection Correlation Analysis is performed using Pearson, Spearman and Kendall coefficients which are explained in algorithm 1, algorithm 2, algorithm 3 and algorithm 4.

**Algorithm 1: Pearson Correlation Analysis**

Pearson correlation coefficient $\rho$ is calculated by the formula as given below:

$$\rho = \frac{E[AD] - E[A]E[D]}{\sqrt{E[A^2] - (E[A])^2}\sqrt{E[D^2] - (E[D])^2}}$$

where:

A stands for the Attribute Vector

D stands for the Decision Vector

E[A] stands for the sum of the elements in A

**Algorithm 2: Spearman Correlation Analysis**

Spearman Correlation coefficient $\sigma$ is calculated by the formula mentioned below:

$$\sigma = 1-(6\Sigma d_i^2)/n(n^2 - 1)$$

Where,

$d_i$ stands for the difference between the ranks of variables P and Q

n stands for the sample size

**Algorithm 3: Kendall Correlation Analysis**

Kendall Correlation coefficient $\tau$ is calculated by the formula as given below:

$$\tau = (n_c - n_d)/(1/2n(n-1))$$

Where,

$d_i$ stands for the difference between the ranks of variables P and Q

n stands for the sample size

After doing Pearson Correlation by Algorithm 1, Spearman Correlation using Algorithm 2 and Kendall-rank Correlation by Algorithm 3, we get a list of attributes that satisfy the respective correlation criteria. After obtaining the three individual results which reduces the number of features using Algorithm 4 discussed below:

**Algorithm 4: Attribute Selection after Correlation**

```
procedure ATTRIBUTESELCTION(Dataset)
rows ← nrows(Dataset)
cols ← ncols(Dataset)
pearsonVector ← pearson(Dataset)
spearmanVector ← spearman(Dataset)
kendallVector ← kendall(Dataset)
for each i in 1:cols do
if pearsonVector[i]>0 AND spearmanVector[i]>0
AND kendallVector[i]>0 then
Selection ← true
else
Selection ← false
end if
end for
return dataset[,Selection]
end procedure
```

### 3.2 Binary Classification

The initial data set had 61 attributes. The data set is modified by adding a 62nd attribute which is Boolean, named 'Popular' and 'Unpopular'. This attribute decides the class label of the data set which is based on average of the number of shares which is explained in algorithm 5.

**Algorithm 5: Deciding Class of Articles**

```
procedure POPULARITY(shares)
sum ← 0
for each i in shares do
sum ← sum + i
end for
avg ← sum
length(shares)
for each i in shares do
if i ≥ avg then
popularity ← true
else
popularity ← false
end if
end for
end procedure
```

For binary classification algorithm 6 is performed and flow diagram is shown in figure 2.

**Algorithm 6: Binary Classification**

```
Input: D {dataset};
Output: Label {Popularity Label};
Step1: For each instance in D, do
Find feature vector (V)
Step 2: For each V do
Feature Reduction using Pearson, Spearman and
Kendall coefficient
Step 3: Data classification using Hybrid Classifier
split data in two halves and classify data using SVM
and RF algorithm
Step 4: Determine the total class label
Find
True_positive (TP)
True_negative (TN)
False_positive (FP)
False_negative (FN)
Step 5: Find Performance Parameters
Step 6: Predict Article popularity Class as
if ( class=1) Article= Popular State
else_if(class =0) Article=Unpopular State
end for
```

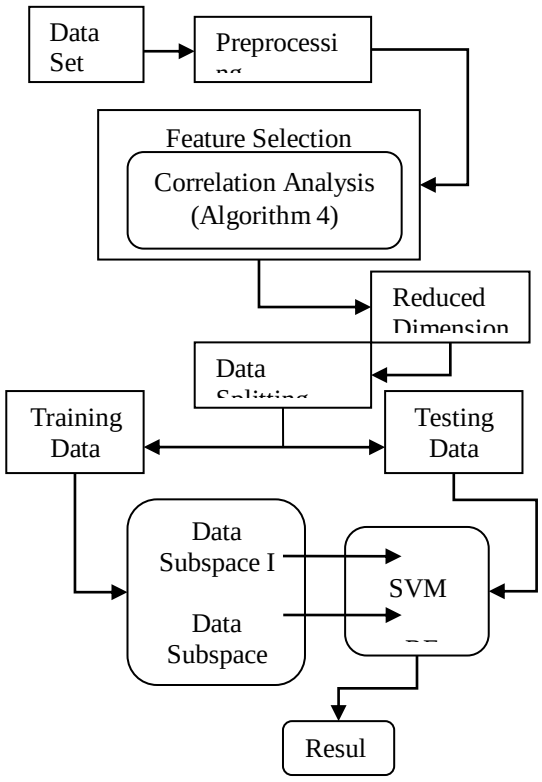


**Fig. 2: Flow diagram of Binary Classification of News Popularity Prediction**

### 3.3 Multi Classification

The initial data set had 61 attributes. The data set is modified by adding a 62nd attribute which is Boolean, named 'Popular' 'Average' and 'Unpopular'. The selected attribute decides the class label of the data set which is based on average and mean of the number of shares which is explained in algorithm 7.

**Algorithm 7: Deciding Multi Class of Articles**

```
procedure POPULARITY(shares)
for each i in shares do
if i<=median(i)
popularity = 0; // popularity = unpopular
else if i>= average(i)
 popularity = 2; // popularity = popular
 else
popularity = 1; // popularity = average
end if end if end for end procedure
```

## 4. IMPLEMENTATION

In order to evaluate the performance of proposed work, the algorithms are executed and their performances are compared.

### 4.1 Data Set Description

The dataset is taken from UCI machine learning repository [13]. This dataset is collected from popular news web site known as Mashable.com. It is preprocessed and donated on this UCI repository [2]. Total 61 attributes are extracted from 39,797 news articles and these attributes describe different features of every article. These news articles are collected during 2 years of period, from January 7 2013 to January 7 2015.

### 4.2 Result Analysis

A news article is popular or unpopular is predicted based on last column of dataset known as 'number of shares' of news article on social media. Threshold value is calculated on 'number of shares' attribute using algorithm 5. The entire dataset is split into training and testing set.
In this work, ten prediction algorithms results are analyzed in order to find which algorithm will give us the maximum prediction rate for reduced attributes.
In Table I and figure 3 result analysis is represented in the form of the percentage values of accuracy for all the prediction algorithms i.e. classifiers for binary classifier. The table data shows that proposed algorithm gives best prediction value for the reduced Attribute Set.

**Table I: Comparative Analysis for Binary Classification**

| Techniques | Accuracy (in %) |
|---|---|
| Hybrid SVM-RF | 99.64 |
| KNN (k=8) | 82.89 |
| AdaBoost | 82.52 |
| LPBoost | 82.49 |
| Naíve bayes [Choudhary et al.] | 93.46 |
| Neural networks [Choudhary et al.] | 91.89 |
| C4.5 tree [Choudhary et al.] | 88.82 |
| Random forest [Choudhary et al.] | 79.63 |
| C5.0 tree [Choudhary et al.] | 79.92 |
| MLR [Choudhary et al.] | 61.11 |

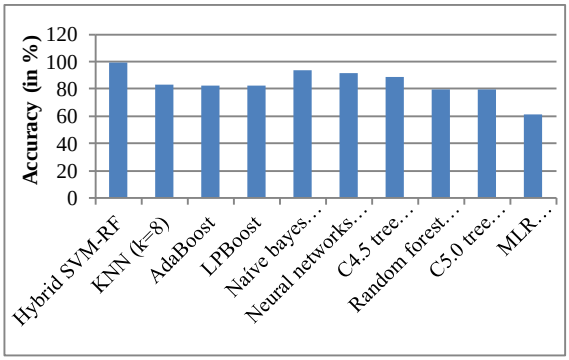


**Fig. 3: Comparative Analysis for Binary Classification**

For multiclassification popularity prediction result analysis is shown in Table II and figure 4. Threshold value is calculated on 'number of shares' attribute using algorithm 7. The entire dataset is split into training and testing set in the ratio of 60:40, 70:30, 75:25 and 80:20 respectively. Further three classifiers, i.e. hybrid SVMRF, Naïve Bayes and KNN, are used to give the maximum prediction rate for reduced attributes.

**Table II: Comparative Analysis for Multi Classification**

| | Accuracy ( in %) | | | |
|---|---|---|---|---|
| | 60:40 Ratio | 70:30 Ratio | 75:25 Ratio | 80:20 Ratio |
| Hybrid SVM-RF | 71.12 | 72.86 | 73.06 | 73.58 |
| Naïve Bayes | 65.98 | 67.04 | 66.64 | 56.20 |
| KNN | 62.12 | 62.63 | 62.78 | 63.26 |

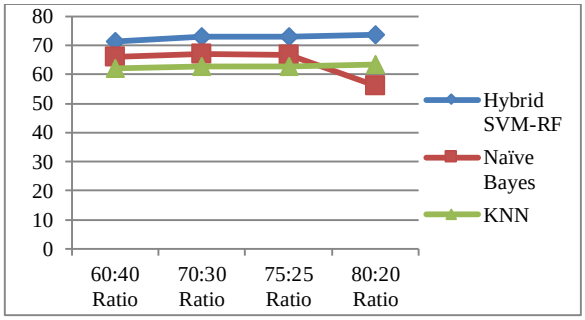


**Fig. 4: Comparative Analysis for Multi Classification**

## 5. CONCLUSION

News popularity is the maximum growth of attention given for particular news article. Online news popularity depends upon various factors such as number of shares on social media, number of comments by visitors, number of likes etc. So it is necessary to build an automated decision support system to predict the popularity of news as it will help in business intelligence too. The work presented in this research intends to find the best model to predict the popularity of online news by using machine learning methods.

After applying Pearson's, Kendall's and Spearman's Correlation Coefficients, out of 61 attributes, attributes are reduced using correlation coefficients techniques. So reduced features are taken into consideration for prediction of popularity. In this work, performance evaluation metrics such as accuracy values are increased and given better performance of classification that is compared with existing research's implemented methodology. The machine learning methods like LPBoost, AdaBoost, Naïve Bayes and KNN is analyzed. From the experimental results, it is observed that proposed algorithm gives the better prediction of about 99.6% for binary classification.

For multiclassification, proposed algorithm, KNN and Naïve Bayes are used to analyse the performance and proposed algorithm achieved the highest accuracy of about 73%.

The evaluation measures such as F-measure and accuracy for popularity prediction can be further improved by applying the natural language processing tools and techniques to understand the semantics of the text. Further the research methods can be applied to find the domain of news popularity too.

## REFERENCES

[1] Ilias N. Lymperopoulos, "Predicting the popularity growth of online content", Elsevier, vol. 369, pp. 585-613, 10 November 2016.

[2] Kelwin Fernandes, Pedro Vinagre, Paulo Cortez, "A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News", Springer, EPIA 2015, pp. 535-546, 2015.

[3] He Ren, Quan Yang, "Predicting and Evaluating the Popularity of Online News", Standford University Machine Learning Report.

[4] Bandari Roja, Sitaram Asur, and Bernardo A. Huberman. "The pulse of news in social media: Forecasting popularity." arXiv preprint arXiv:1202.0332, 2012.

[5] Ioannis Arapakis, B. Barla Cambazoglu, and Mounia Lalmas, "On the Feasibility of Predicting News Popularity at Cold Start", Springer, pp. 290-299, 2014.

[6] R. Shreyas, D.M Akshata, B.S Mahanand, B. Shagun, C.M Abhishek, "Predicting Popularity of Online Articles using Random Forest Regression", International Conference on Cognitive Computing and Information Processing, IEEE, 2016

[7] Alexandru Tatar, Marcelo Dias de Amorim, Serge Fdida and Panayotis Antoniadis, "A survey on predicting the popularity of web content", Journal of Internet Services and Applications 2014, A Springer Open Journal, pp. 1-20, 2014.

[8] Alexandru Tatar, Panayotis Antoniadis, Marcelo Dias de Amorim, Serge Fdida, "From Popularity Prediction to Ranking Online News", HAL, pp. 1-14, 2014.

[9] Ioannis Arapakis, B. Barla Cambazoglu, and Mounia Lalmas, "On the Feasibility of Predicting News Popularity at Cold Start", Springer, pp. 290-299, 2014.

[10] Ren He and Quan Yang."Predicting and Evaluating the Popularity of Online News."

[11] Swati Choudhary, Angkirat Singh Sandhu and Tribikram Pradhan, "Genetic Algorithm Based Correlation Enhanced Prediction of Online News Popularity" Computational Intelligence in Data Mining, Advances in Intelligent Systems and Computing, Springer, 2017, pp.133-144.

[12] UCI Machine Learning Database, https://archive.ics.uci.edu/ml/datasets/Online+News+Popularity, May 2015.